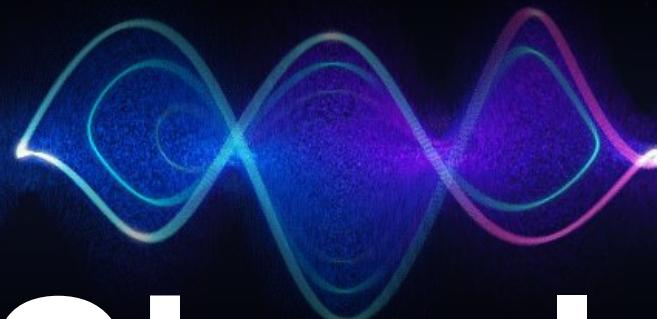


# Rap Simulator

머신러닝을 활용한 입문자 랩 스타일 추천



# Rap Simulator

힙합 음악의 주요 요소, 반복되는 비트를 배경음으로 박자에 맞춰 가사를 내뱉는 것

모형을 이용한 모의실험 도구

# INDEX



## PART1

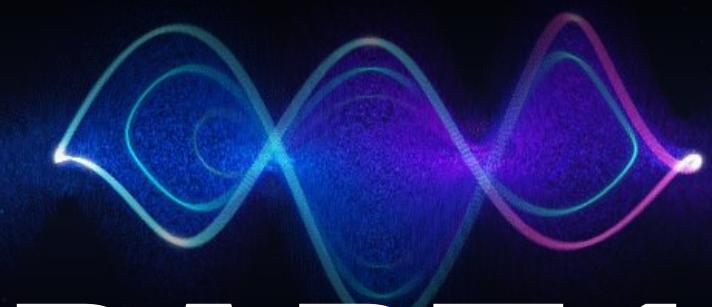
- 프로젝트 개요
- 프로젝트 팀 구성 및 역할

## PART2

- 프로젝트 수행 절차 및 방법
- 프로젝트 수행 경과

## PART3

- 자체 평가 의견



# PART 1

프로젝트 개요  
프로젝트 팀 구성 및 역할

내가 참고하면 좋을 래퍼는 누구일까?

곡이 너무 많은데 어떤 곡으로 연습해야 할까?

랩에 흥미를 느꼈지만 어디서부터 시작해야 할지 막막한 사람들  
⇒ 목소리, 억양, 플로우 분석으로 참고 래퍼와 곡을 추천하는

## 'Rap Simulator'

- ✓ 아티스트 매칭
- ✓ 맞춤 연습곡 추천
- ✓ 레코딩 기록 및 공유

# 아티스트 선정

프로젝트 개요

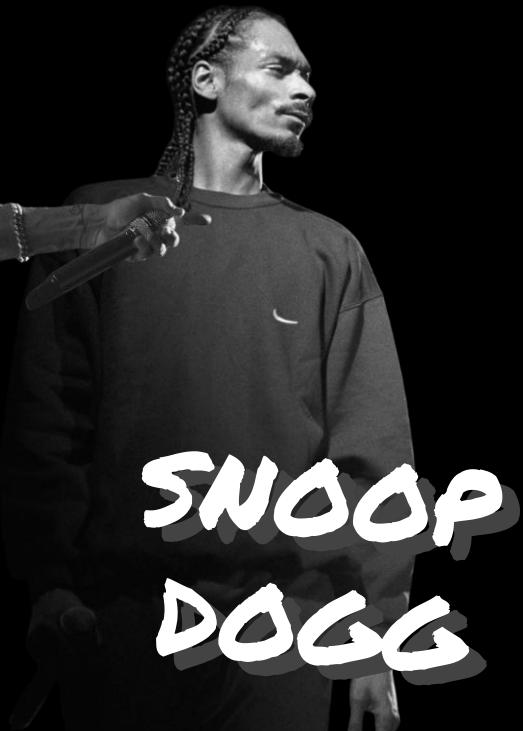
111



JAY Z



EMINEM



SNOOP  
DOGG

## 1 인지도

세계적으로 유명해 추천 결과의 납득성이 높고, 실제로 많은 지망생이 보고 따라하는 래퍼

## 2 변인 통제

영어권 남성 래퍼 -> 언어, 성별 변인을 통제하여 모델의 운율과 음색 구분 능력을 검증

## 3 데이터 활용성

장기간 활동하여 라이브, 공개 음원 등 수집 가능한 데이터가 비교적 풍부함

## 4 랩핑 스타일 차이

힙합 분야에서 자신만의 독자적인 스타일을 구축한 래퍼들을 선정

# 프로젝트 목표

## 1 모델 학습

아티스트 음성 기반으로 목소리,  
억양, 피치, 플로우 등을 학습

프로젝트 개요

1

일반인의 음성을 분석하여  
유사성 높은 아티스트 매칭

## 2 유사도 계산

# 프로젝트 의의

## 1 접근성 상승

랩의 언더그라운드적 특성 | 높은 진입 장벽 탓에 대개 독학으로 시작

높은 레슨비 | 전문 래퍼의 개인 레슨 외 대안이 없어 비용적 부담

Rap Simulator | 모바일 어플리케이션으로 손 쉽게 랩 시작

프로젝트 개요

1

추천→연습→피드백→기록 | 사용자 경험

사용자가 자신의 랩 특성을 파악할 수 있도록 정보 제공 | Rap Simulator

내가 무엇을 알고, 무엇을 모르는지 인지 | Metacognition

## 2 메타인지 향상

역할 분담

SUNG JAE LEE

팀장 | 발표 | 모델 탐색

JIN WOOK KIM

데이터 수집 및 정제 | 영상 편집

DONG KYUN RYU

목업 제작 | 모델 탐색

프로젝트 팀 구성 및 역할

1W

YOUNG RAN KO

데이터 분석 | 모델 탐색

JI EON PARK

발표 자료 작성 | 모델 탐색

WON JAE LEE

데이터 수집 및 정제 | 데이터 분석



# PART 2

프로젝트 수행 절차 및 방법  
프로젝트 수행 경과

# 프로젝트 캘린더

프로젝트 수행 절차 및 방법

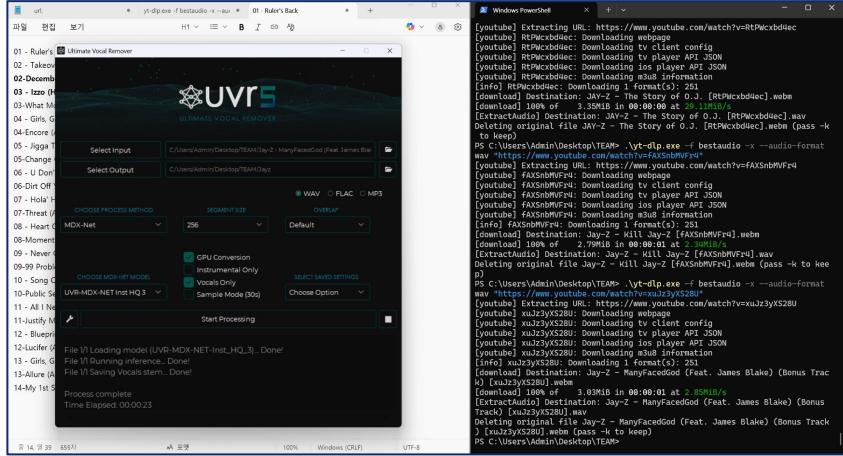


2025년 7월 23일 ~ 8월 14일

주제 선정	7/23 - 7/28
데이터 수집/정제	7/29 - 8/1
모델 선정	8/1 - 8/4
데이터 분석 및 모델링	8/1 - 8/11
목업 제작	8/3 - 8/6
발표 자료 제작	8/4 - 8/11
발표/질의응답 준비	8/8 - 8/14

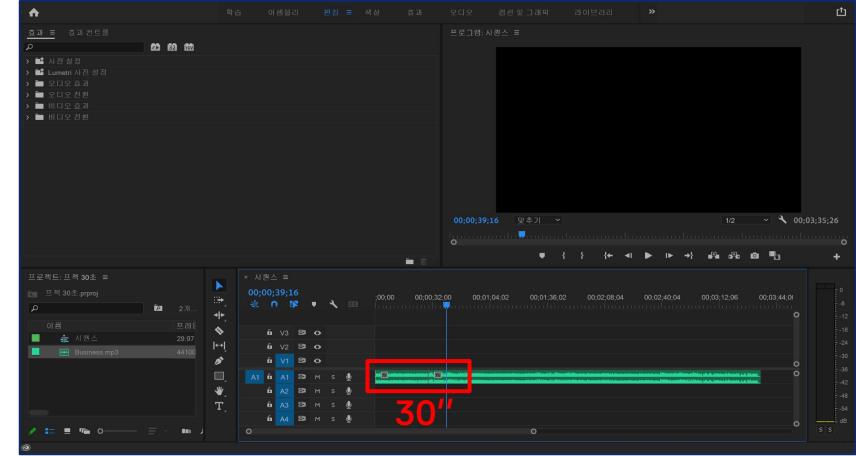
# 데이터 수집/정제

## 프로젝트 수행 경과



### [반주-보컬 분리]

- yt-dlp 명령어 프로그램으로 WAV 파일 생성
- UVR5 AI 프로그램으로 보컬만 추출



### [음성 30초 분할]

- 보컬 추출 음원에서 피처링을 제외한 타겟 가수의 음성만 채택
- 30초 분할로 2차 데이터 가공

# 데이터 수집/정제

## 프로젝트 수행 경과

유형	사람	수정 날짜	출처	유형	사람	수정 날짜	출처
이름	이름	이름		이름	이름	이름	
Streets Is Watching (Vocals).2.mp3	You Don_t Know.1.mp3	You Thought(Vocals).1.mp3		You Thought(Vocals).1.mp3	Won_t Back Down.3.mp3	Y_All Gone Miss Me(Vocals).3.mp3	
Streets Is Watching (Vocals).1.mp3	Won_t Back Down.2.mp3	Y_All Gone Miss Me(Vocals).2.mp3		Won_t Back Down.2.mp3	Won_t Back Down.1.mp3	Y_All Gone Miss Me(Vocals).1.mp3	
So Ghetto (Vocals).5.mp3	Without Me.3.mp3	Wrong Idea(Vocals).1.mp3		Without Me.3.mp3	Without Me.2.mp3	Who Am I (Vocals).1.mp3	
So Ghetto (Vocals).4.mp3	Without Me.2.mp3	Who Am I (Vocals).2.mp3		Without Me.1.mp3	Vato, (Dirty) (Vocals).1.mp3	Vato, (Dirty) (Vocals).2.mp3	
So Ghetto (Vocals).3.mp3	White America.4.mp3	Vapors(Vocals).4.mp3		White America.4.mp3	Vapors(Vocals).3.mp3	Vapors(Vocals).2.mp3	
So Ghetto (Vocals).2.mp3	White America.3.mp3	Vapors(Vocals).1.mp3		White America.3.mp3	When Im Gone.6.mp3	Up Jump Tha Boogie(Vocals).1.mp3	
So Ghetto (Vocals).1.mp3	White America.2.mp3	Vapors(Vocals).2.mp3		White America.2.mp3	When Im Gone.5.mp3	Trust Me (Vocals).1.mp3	
Snoopy Track (Vocals).2.mp3	White America.1.mp3	Vapors(Vocals).3.mp3		White America.1.mp3	When Im Gone.4.mp3	The Shiznit (Vocals).3.mp3	
Snoopy Track (Vocals).1.mp3	When Im Gone.6.mp3	The Shiznit (Vocals).2.mp3		When Im Gone.6.mp3	When Im Gone.5.mp3	The Shiznit (Vocals).1.mp3	
Ride Or Die (Vocals).6.mp3	When Im Gone.4.mp3	The Shiznit (Vocals).3.mp3		When Im Gone.4.mp3	When Im Gone.3.mp3	The Shiznit (Vocals).2.mp3	
Ride Or Die (Vocals).5.mp3	When Im Gone.2.mp3	The Shiznit (Vocals).1.mp3		When Im Gone.2.mp3	When Im Gone.1.mp3	The Shiznit (Vocals).0.mp3	
Ride Or Die (Vocals).4.mp3	When Im Gone.0.mp3	The Shiznit (Vocals).0.mp3		When Im Gone.0.mp3	Ride Or Die (Vocals).3.mp3	The Shiznit (Vocals).0.mp3	
Ride Or Die (Vocals).3.mp3	Ride Or Die (Vocals).2.mp3	The Shiznit (Vocals).0.mp3		Ride Or Die (Vocals).2.mp3	Ride Or Die (Vocals).1.mp3	The Shiznit (Vocals).0.mp3	
Ride Or Die (Vocals).2.mp3	Ride Or Die (Vocals).0.mp3	The Shiznit (Vocals).0.mp3		Ride Or Die (Vocals).0.mp3	Ride Or Die (Vocals).0.mp3	The Shiznit (Vocals).0.mp3	

### [ 훈련용 데이터 ]

- 스눕독, 제이지, 에미넴의 음원 30초 클립으로 구성
- 100개씩 총 300개 사용

유형	사람	수정 날짜	출처
제이지 Holy Graii 샘플 30초.mp3	제이지 라이브.mp3		
제이지 Dead Presidents 샘플 30초.mp3	에미넴 라이브.mp3		
제이지 Big Pimpin 샘플 30초.mp3	스눕 독 진과 주스 라이브.mp3		
에미넴 without me 샘플 30초.mp3			
에미넴 stan 샘플 30초.mp3			
에미넴 rap god 샘플 30초.mp3			
스눕독 the next episode 샘플 30초.mp3			
스눕독 nuthin but a g thang 샘플 30초.mp3			
스눕독 gin & juice 샘플 30초.mp3			

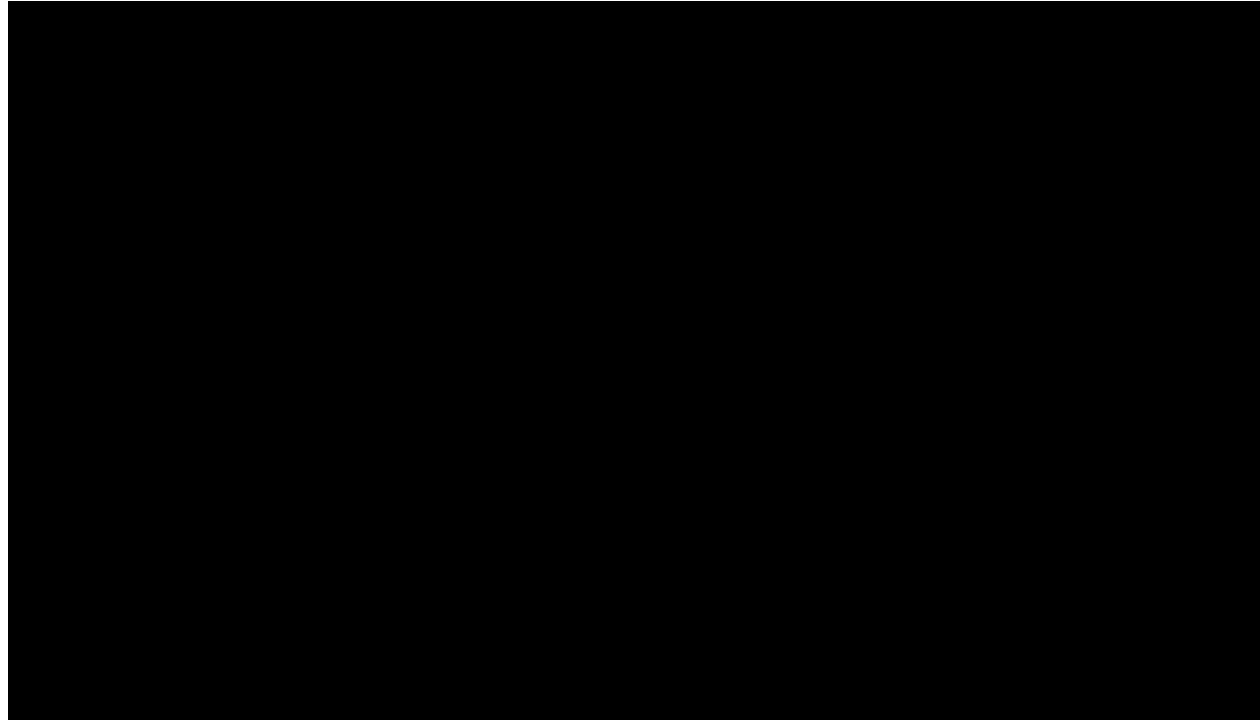
### [ 테스트용 데이터 ]

- 훈련 데이터에 포함되지 않은 음원으로 구성
- 일반인 커버곡 9곡과 래퍼 라이브 음원 3곡, 총 12곡 사용

학습한 모델이 일반인의 랩 특성을 분석해 **아티스트와의 유사도를 판별**할 수 있는지 확인하고자 함

# 테스트셋 데이터 예시

프로젝트 수행 경과



# 머신러닝

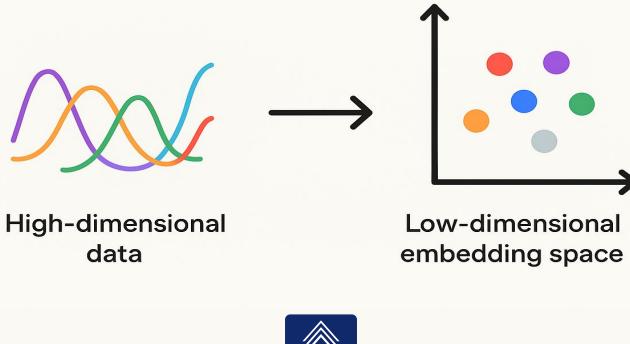
프로젝트 수행 경과



#	ML Classifier	Result
1	XGBoost	Accuracy=0.72
2	RandomForest, SVM	Accuracy=0.69
3	Gaussian Naive-Bayes	Accuracy=0.73
결론	<ul style="list-style-type: none"><li>적은 데이터 수로 인한 과적합 문제</li><li>GridSearchCV 등으로 하이퍼파라미터 튜닝을 거듭하였으나, 머신러닝 분류기의 성능 개선에 한계 존재</li><li>오디오 데이터를 더욱 잘 학습할 수 있는 딥러닝 모델 탐색</li></ul>	



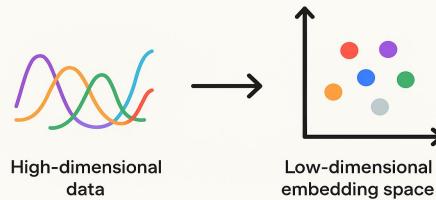
## Embedding



- 음성, 텍스트, 이미지 등 비정형 데이터를 기계가 처리할 수 있는 숫자 벡터로 변환하는 과정이나 그 결과 예: 원핫인코딩
- 오디오의 **파형(wave form)**을 입력 데이터로 받아 처리하는 transformer 기반 wav2vec2.0와 HuBERT 모델



## Embedding



### [ wav2vec2.0 ]

- 맥락 기반 추론
- 자기주도학습  
목소리를 잘게 쪼개서 패턴을 추출한 뒤 라벨 없이 직접 학습
- 추출한 소리의 특징을 바탕으로 전체 음성의 흐름과 맥락을 이해하여 다음 소리를 예측

### [ HuBERT ]

- 클러스터 기반 추론
- 비슷한 소리를 클러스터로 묶어 임시 라벨을 만든 뒤 라벨을 예측
- wav2vec2보다 발음·역양·리듬 같은 말소리 구조를 깊게 이해  
⇒ 래퍼별 목소리 차이, 감정, 말투를 더 세밀하게 구분

# 데이터 전처리

프로젝트 수행 경과



```
49 # 전처리: 리샘플링(sr), 무음 제거, 30초(dur) 자르기, 증강 적용
50 def preprocess_task(args):
51     path, augment, artist = args
52     y, _ = librosa.load(path, sr=SR)
53     intervals = librosa.effects.split(y, top_db=TOP_DB)
54     y = np.concatenate([y[s:e] for s,e in intervals])
55     L = int(DUR * SR)
56     chunk = y[:L]
57     if len(chunk) < L:
58         chunk = np.pad(chunk, (0, L-len(chunk)))
59     chunk = librosa.util.normalize(chunk)
60     if augment:
61         chunk = augmenter(samples=chunk.astype(np.float32), sample_rate=SR)
62     return chunk, artist
```

	embedding	artist	操作
0	[tensor(-0.0294), tensor(0.0041), tensor(-0.08...	Snoop Dogg	수정
1	[tensor(-0.0254), tensor(0.0166), tensor(-0.11...	Snoop Dogg	수정
2	[tensor(-0.0444), tensor(0.0227), tensor(-0.09...	Eminem	
3	[tensor(-0.0228), tensor(0.0108), tensor(-0.06...	Eminem	
4	[tensor(-0.0210), tensor(0.0079), tensor(-0.10...	Eminem	
5	[tensor(-0.0198), tensor(0.0007), tensor(-0.11...	Eminem	
6	[tensor(-0.0309), tensor(0.0080), tensor(-0.09...	Eminem	
7	[tensor(-0.0332), tensor(0.0110), tensor(-0.09...	Snoop Dogg	
8	[tensor(-0.0224), tensor(0.0022), tensor(-0.10...	Snoop Dogg	
9	[tensor(-0.0252), tensor(0.0074), tensor(-0.10...	Snoop Dogg	

## [ sampling rate | 16,000 Hz(헤르츠) ]

- 소리를 디지털 신호로 변환할 때, 초당 샘플링 횟수를 표시
- ⇒ 일반적으로 CD 및 스트리밍 음질은 44.1Hz를 사용하나, 학습 정확도를 높이기 위해 초당 16,000Hz를 샘플링

## [ duration | 30 sec ]

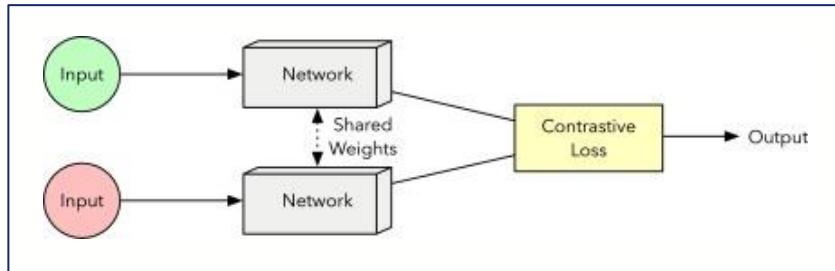
- 데이터를 30초씩 분할

## [ augmentation | p=0 ]

- 부족한 데이터를 보강하기 위해 증강 시도:  
노이즈 추가, 피치 변화, 타임 스트레치 등
- ⇒ 결과적으로 증강하지 않았을 때 고성능 확인,  
증강 확률을 0으로 설정

# 딥 러닝 | 모델

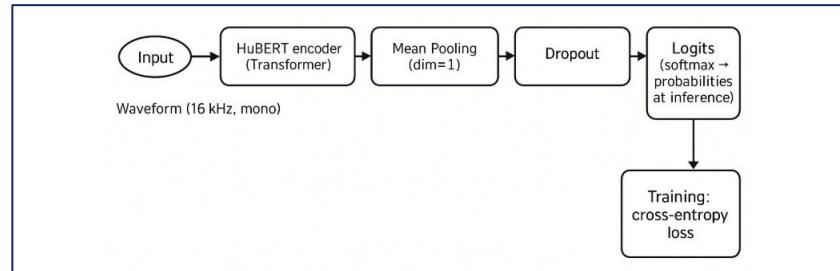
프로젝트 수행 경과



## [ 샘 네트워크 (SiameseNetwork) ]

- 쌍둥이 네트워크
- 한 쌍의 신경망이 가중치를 공유하여, 데이터 쌍의 유사도 식별
- 데이터 셋이 작아도 사용 가능
- nn.Sequential 모듈로 Linear 층과 ReLU 층 구현
  - ➔ positive일 경우, label = 1
  - ➔ negative일 경우, label = 0

※ 손실함수: ContrastiveLoss



## [ 자동오디오분류모델 (AutoModelforAudioClassification) ]

- 사람 목소리를 숫자 벡터로 변환한 뒤,  
각 클래스(예: 래퍼)에 속할 확률을 계산해 분류하는 모델
- 사전학습된 HuBERT 인코더가 특징을 추출하고,  
Linear 층이 최종 클래스를 예측

➔ 입력 오디오를 HuBERT 인코더로 특성(Feature) 벡터로 변환  
➔ 마지막 출력층(Linear)이 각 클래스(래퍼)에 대한 확률 값 산출

# 딥 러닝 | 모델

프로젝트 수행 경과

[ 샘 네트워크 (SiameseNetwork) ]

```
1 class SiameseEmbNet(nn.Module):
2     def __init__(self, dim):
3         super().__init__()
4         self.fc = nn.Sequential(
5             nn.Linear(dim, 256),
6             nn.ReLU(),
7             nn.Linear(256, 128)
8         )
9     def forward(self, x):
10        return self.fc(x)
```

[ 자동오디오분류모델 (AutoModelforAudioClassification) ]

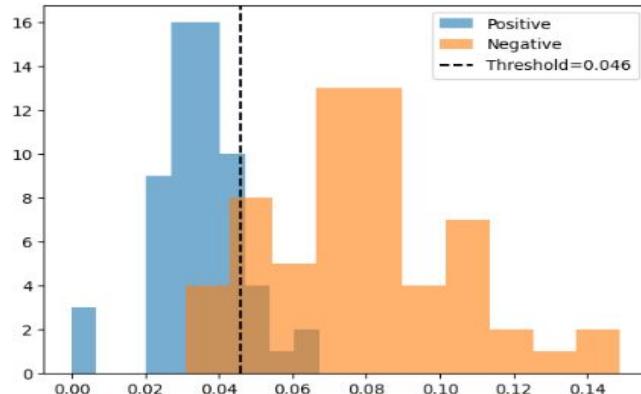
```
1 class HubertForSequenceClassification(HubertPreTrainedModel):
2     def __init__(self, config):
3         super().__init__(config)
4         self.hubert = HubertModel(config)
5         self.dropout = nn.Dropout(config.final_dropout)
6         self.classifier = nn.Linear(
7             config.hidden_size, config.num_labels
8         )
9     def forward(self, input_values=None):
10        outputs = self.hubert(input_values)
11        hidden_states = outputs[0]
12        pooled = hidden_states.mean(dim=1)
13        pooled = self.dropout(pooled)
14        logits = self.classifier(pooled)
15        return logits
```

# 최종 결과 | Siamese Network

프로젝트 수행 경과



```
==== Fold 1 ====
Fold1 Epoch 1 Loss: 0.0025
Fold1 Epoch 2 Loss: 0.0020
Fold1 Epoch 3 Loss: 0.0018
Fold1 Epoch 4 Loss: 0.0017
Fold1 Epoch 5 Loss: 0.0015
Fold1 Epoch 6 Loss: 0.0013
Fold1 Epoch 7 Loss: 0.0012
```



Fold1 AUC: 0.952, Accuracy: 0.900

Fold1 Confusion Matrix:

[[55 4]	[ 8 53]]
---------	----------

## [ Validation 결과 ]

- validation 정확도: 0.82~0.90
- Epoch이 반복될 수록 손실이 감소, 학습 능률 상승
- 100건 중 약 87건을 바르게 분류

# 최종 결과 | Siamese Network

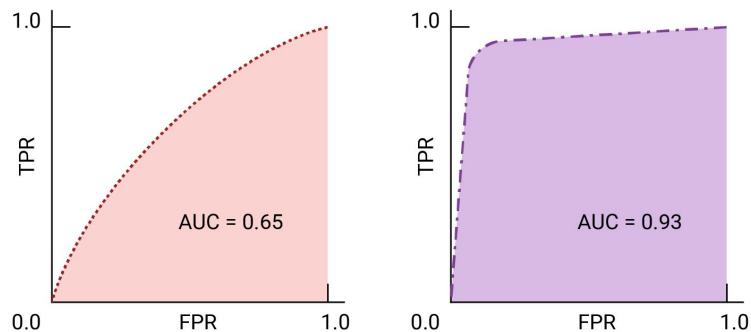
프로젝트 수행 경과



## ==== Cross-Validation Summary ====

Mean AUC:  $0.932 \pm 0.032$

Mean Accuracy:  $0.865 \pm 0.044$



## [ Validation 결과 ]

- validation 정확도:  $0.82\sim0.90$
- Epoch이 반복될 수록 손실이 감소, 학습 능률 상승
- 100건 중 약 87건을 바르게 분류

### ➔ AUC란:

ROC Curve의 아래 면적  
1에 근접할 수록 우수

# 최종 결과 | Siamese Network

프로젝트 수행 경과



Test Accuracy: 0.889

Test Confusion Matrix:

	Eminem	Jay Z	Snoop Dogg
Eminem	3	0	0
Jay Z	0	2	1
Snoop Dogg	0	0	3

File: Snoop Dogg - the next episode 샘플 30초

Eminem : 99.47%

Jay Z : 99.94%

Snoop Dogg: 99.95%

→ Predicted: Snoop Dogg, Confidence: 99.95%

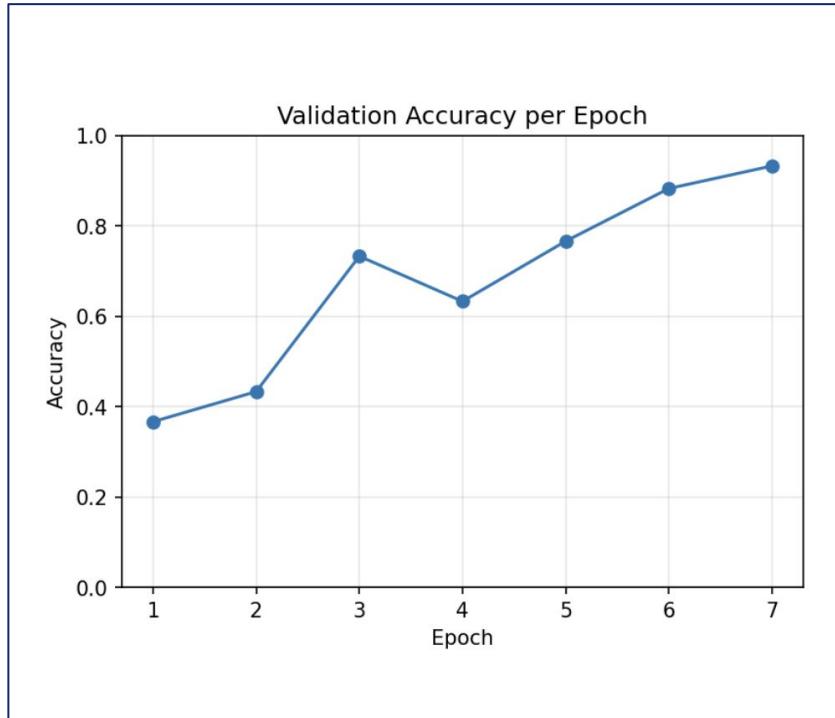
## [ Test 결과 ]

- 일반인의 랩 커버 음성인 test set의 정확도는 0.89로 원곡자를 잘 판별

⇒ 한계: Softmax 및 코사인 유사도를 활용한 거리 기반 유사도 계산 시에는 타가수와 큰 차이가 나지 않음

# 최종 결과 | HuBERT + Audio Classifier

프로젝트 수행 경과

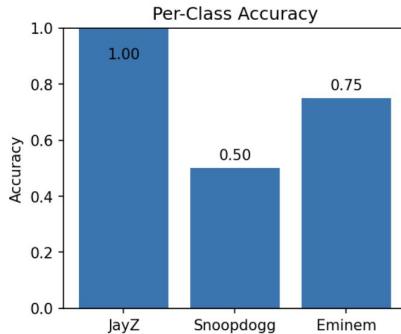


## [ Validation 결과 ]

- Epoch 1~2  
정확도가 약 43.3%로, 랜덤 추측 수준
- Epoch 3  
모델이 특징을 학습하며 성능이 급상승
- Epoch 6  
정확도 88.3% (향상)
- Epoch 7  
정확도 93.3% (향상)

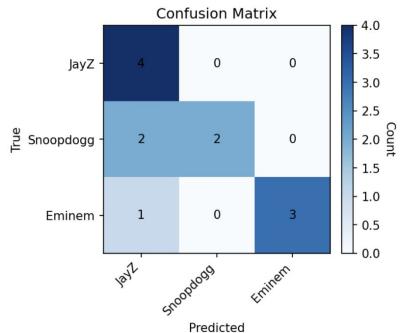
# 최종 결과 | HuBERT + Audio Classifier

프로젝트 수행 경과



## [ Test 결과 ]

- test set에서 정확도는 0.75  
JAY Z: 4/4 정답 → 안정적으로 분류  
Snoop Dogg: 2/4 정답, 2건이 JAY Z로 오분류  
Eminem: 3/4 정답. 1건이 JAY Z로 혼동

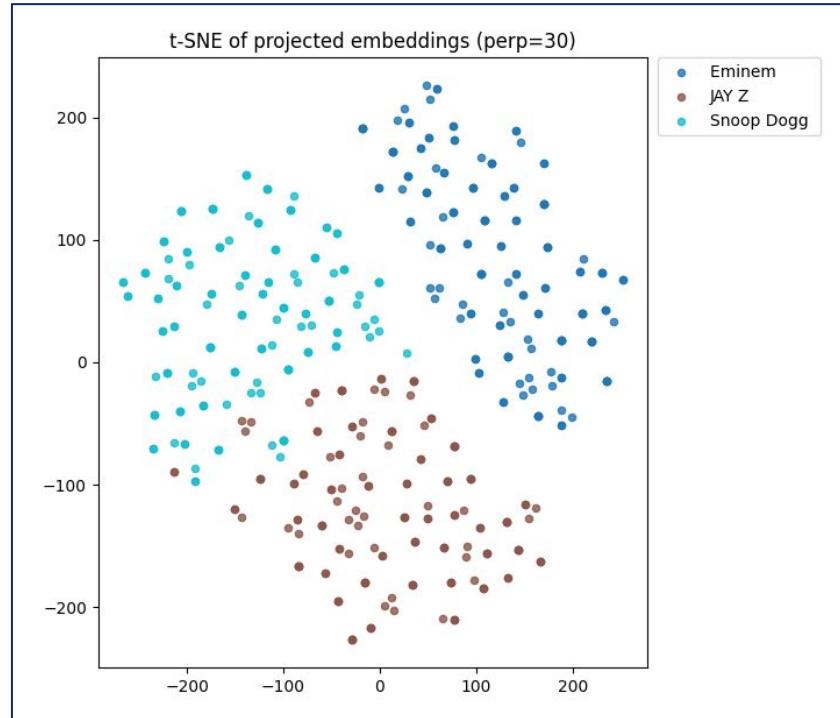


# 결론

## 프로젝트 수행 경과

### [ 모델 선정 ]

- 분류 정확도는 삼 네트워크 모델이 더 높음
  - 차원축소 기법으로 2차원 평면 상에 군집을 표현한 결과,  
세 군집이 비교적 잘 구분
- ⇒ 한계: 특정 아티스트를 추천해 주기에는  
군집 간 거리가 가까워,  
유의미한 차이를 제공하지 못함



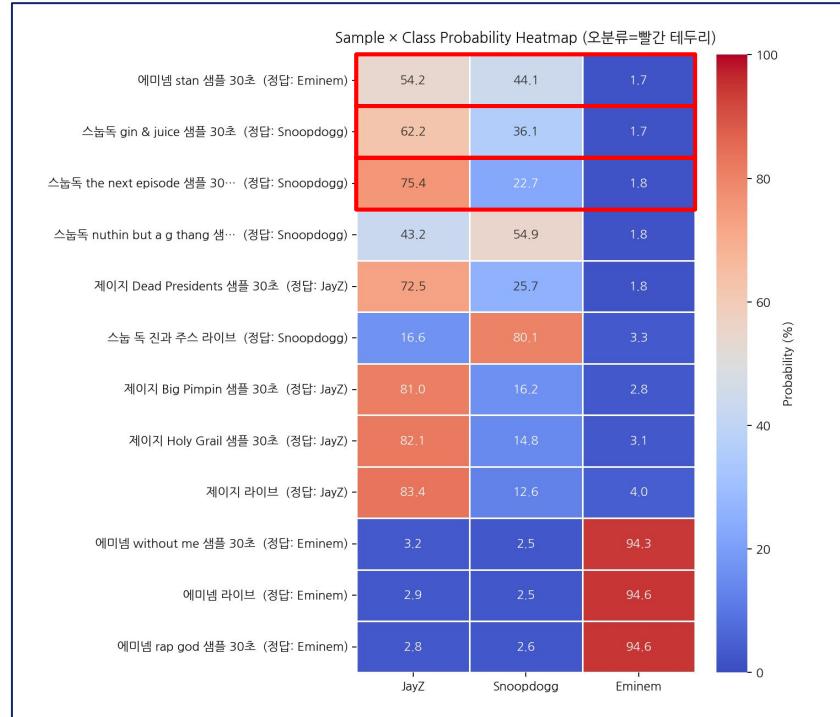
# 결론

## 프로젝트 수행 경과

### [ 모델 선정 ]

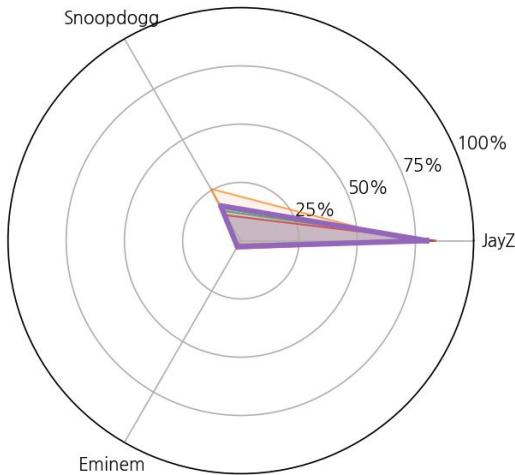
- 본래 목적인 유사도 판별에는 **HuBERT** 분류 모델이 적합
- 유사도는 확률 기반으로 계산하였으며,  
각 샘플에 대해 아티스트 유사도가 뚜렷

⇒ 결과: 최종 모델로 **HuBERT + Audio Classifier** 선정

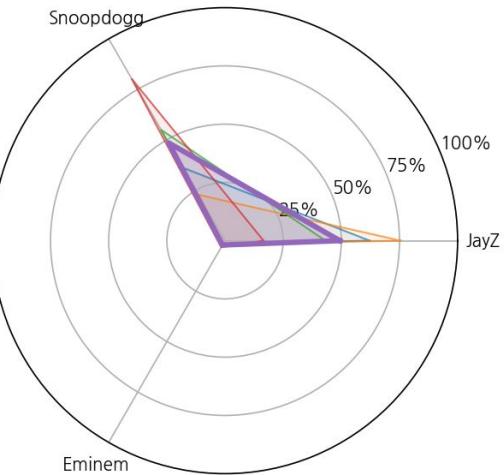




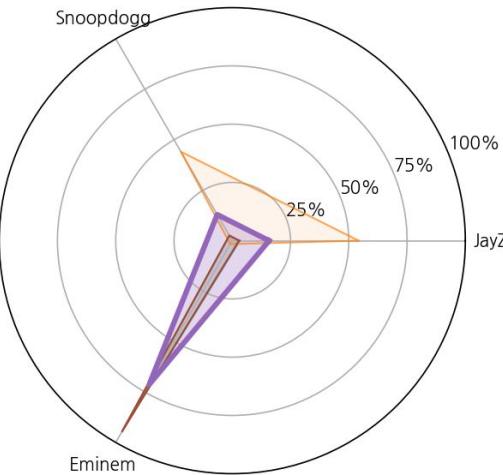
JayZ — Overlay Radar (n=4)

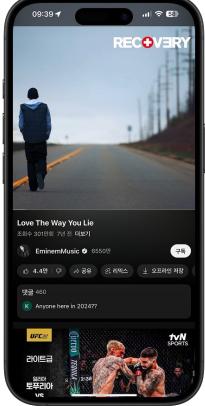
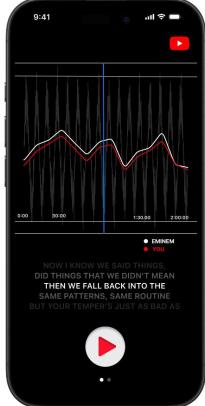
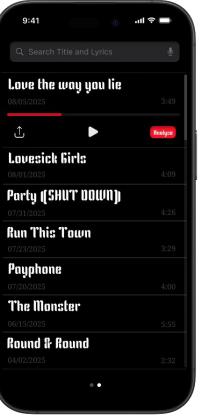
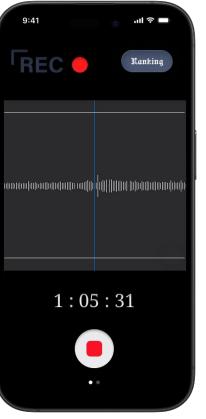
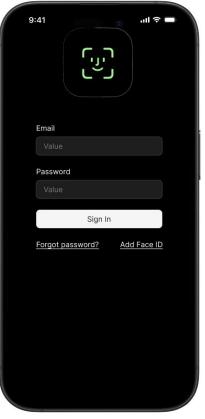


Snoopdogg — Overlay Radar (n=4)



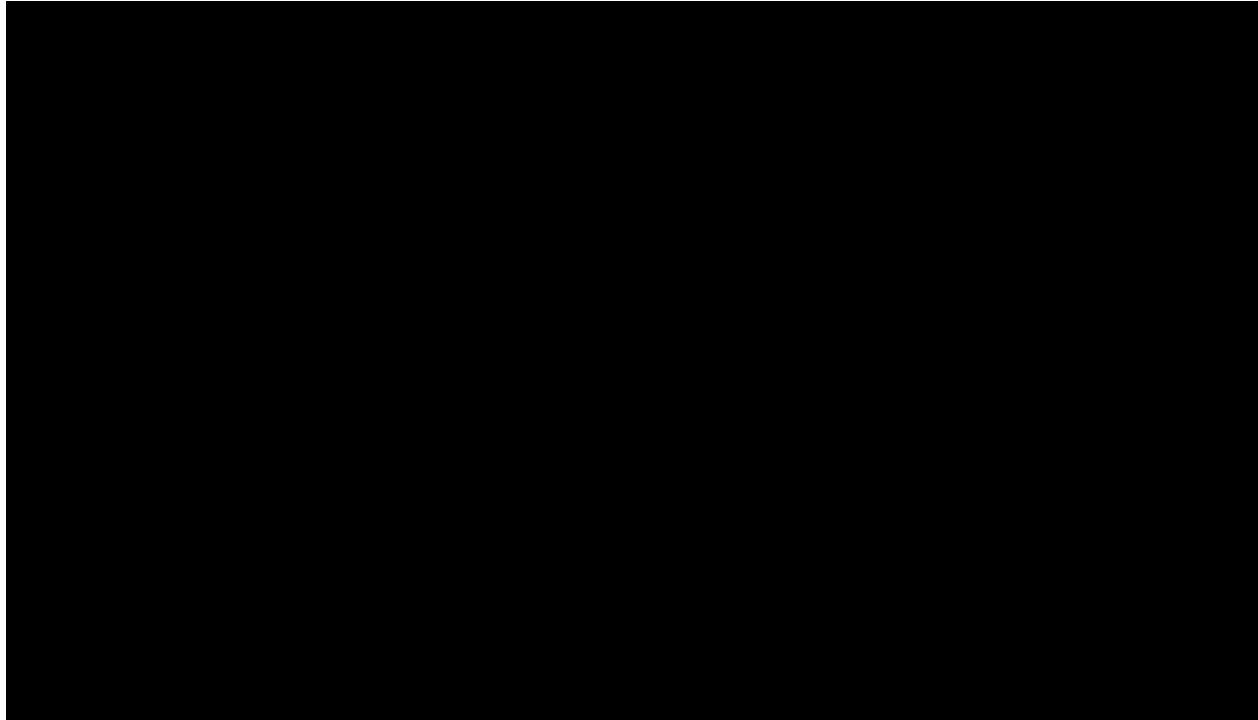
Eminem — Overlay Radar (n=4)





# 목업 영상

프로젝트 수행 경과



# 인사이트 공유

## 프로젝트 수행 경과

### 모델의 한계

#### 1. 데이터 문제

데이터 **부족**으로 인한 과적합 문제

음원에서 타겟 가수의 음성만 완벽하게 분리 불가  
(효과음, 코러스 등 **노이즈** 존재)

mp3에서 **압축된 소리**는 wav로 변환하더라도 복구할 수 없음

#### 2. 일반화 어려움

일부 래퍼에 국한되어 **다양한 스타일**을 시험해 보지 못하였음  
→ 학습한 특성에서 이질적인 데이터가 입력되면 분류 성능 저하

테스트셋 일반화가 어려워 **모델 유연성** 하락

한국인과 서양인의 **발성 차이**를 고려하지 않음

### 프로젝트 인사이트

#### 1. 데이터 개선 방안

음성 데이터에도 다양한 특성이 존재하며,  
이를 **전처리 하는 방식에 따라 다른 결과** 가 도출될 수 있음

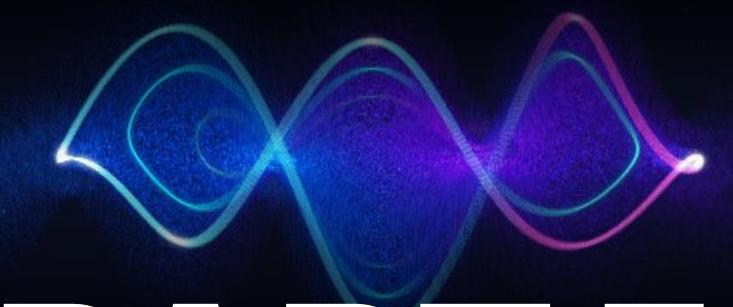
음성 데이터 뿐만 아니라 **감성, 스타일, 가사** 등을 고려해  
데이터셋을 구성하는 접근도 필요

#### 2. 훈련 개선 방안

모델 훈련 시 제이지와 스눕독의 분별이 어려웠는데,  
이는 두 사람 모두 아프리카계 미국인의 억양과 발성을 사용하기  
때문으로 추정

**여성, 아시안 등 다양한 목소리와 언어권 기반 래핑 스타일** 을  
학습시켜 공정성과 모델 범용성 보완 가능

구분이 어려울 정도로 비슷한 음성적 특징을 가진 래퍼들의 경우,  
**추가적 음성 특징 추출과 수치화** 에 대한 고민 필요

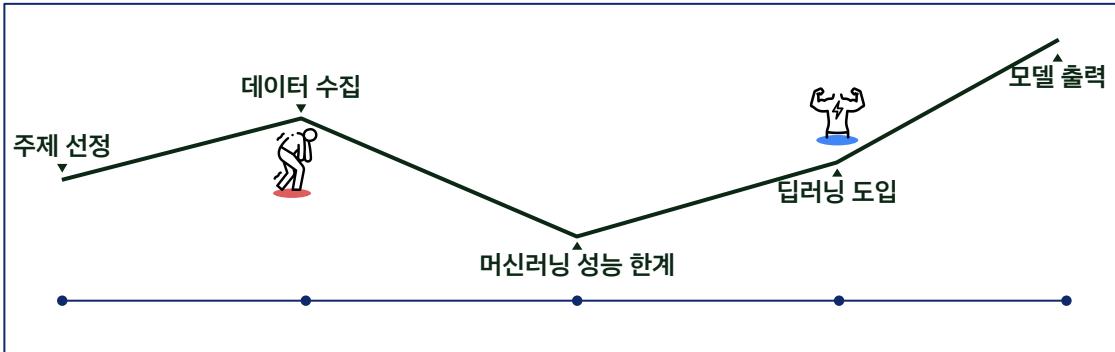


# PART 3

자체 평가 의견

# 성찰 & 차기 계획

자체 평가 의견



## [ 완성도 평가 ]



- 대량의 데이터 수집에 큰 어려움이 있었으나, 구동 가능한 프로토타입을 구현

## [ 차기 방향성 제시 ]

- 레코딩 히스토리 그래프 도입
- 랩 스킬 챌린지 프로그램 구현

데이터...데이터...데이터...!!!

스눕 독이 난 참 밍다.

진육이형 고생하셨습니다.

딥러닝 디핑해봤다:)

딥러닝. 딥하고 스파이시하다.

애들아 3주동안 수고했고  
나중에 웃으면서 보자.

# 참조 자료

프로젝트 수행 경과



#	TITLE	LINK
1	AUC 그래프	<a href="https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=ko">https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=ko</a>
2	삼 네트워크 이미지	<a href="https://www.sciencedirect.com/topics/computer-science/siamese-neural-network">https://www.sciencedirect.com/topics/computer-science/siamese-neural-network</a>
3	Eminem Music	The Monster VEVO, Rap God VEVO, Lose Yourself
4	[Figma] Background Eminem pic	<a href="#">Background Eminem pic</a>
5	[Figma] Eminem pic 1	<a href="#">Eminem pic 1</a>
6	[Figma] Eminem pic 2	<a href="#">Eminem pic 2</a>
7	[Figma] SnoopDogg pic	<a href="#">SnoopDogg pic</a>
8	[Figma] Jay Z pic	<a href="#">Jay Z pic</a>
9	[Figma] Jay Park pic	<a href="#">Jay Park pic</a>
10	Fostering Metacognition to Support Student Learning and Performance	<a href="https://doi.org/10.1187/cbe.20-12-0289">https://doi.org/10.1187/cbe.20-12-0289</a>

# 참조 자료

프로젝트 수행 경과



#	TITLE	LINK
11	Rap music and the empowerment of today's youth: Evidence in everyday music listening, music therapy, and commercial rap music	<a href="https://doi.org/10.1007/s10560-012-0285-x">https://doi.org/10.1007/s10560-012-0285-x</a>
12	A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding	<a href="https://doi.org/10.48550/arXiv.2111.02735">https://doi.org/10.48550/arXiv.2111.02735</a>
13	Siamese Network's Performance for Face Recognition	<a href="https://doi.org/10.1109/ICSECC51444.2020.9557529">https://doi.org/10.1109/ICSECC51444.2020.9557529</a>



Peace  
Out!